Review of reaction datasets available

This review aims to present the reaction datasets available in the ORD and give a short description of the chemistry they are studying. Making it easy to estimate the chemical space explored in each of them.

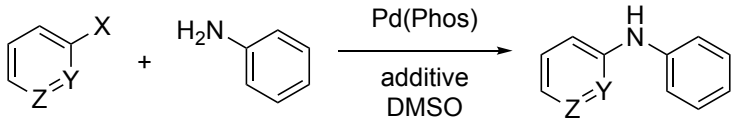
1 HTE - Screenings :

1.1 hte_ahneman :

The Ahneman dataset is a well known HTE dataset.[1]

It is available in the ORD : data/46/ord_dataset-46ff9a32d9e04016b9380b1b1ef949c3.pb.gz

Number of reactions	Number of variables
4312	Ligand : 4, Bases : 3, Additives : 23, Ar_halides : 15

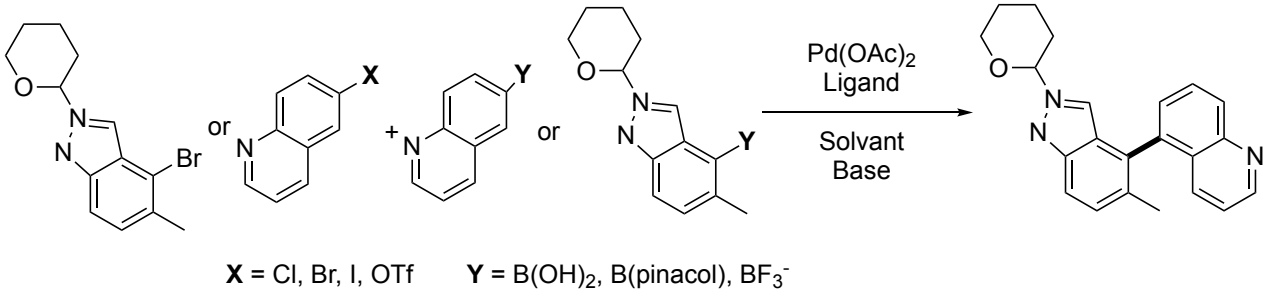


1.2 hte_perera :

The Perera dataset is a well known HTE dataset.[6]

It is available in the ORD : 'data/68/ord_dataset-68cb8b4b2b384e3d85b5b1efae58b203.pb.gz'

Number of reactions	Number of variables
5760	Ar_halide :5, solvents :4, Ligand_in_solvent :(12x2), Base_in_solvent :(8x4), Boronate_in_solvent (8x3)

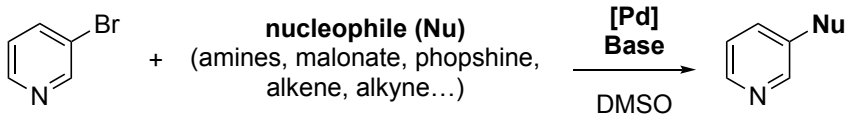


1.3 hte_santanilla :

The Santanilla dataset is a HTE dataset of a Pd cross coupling.[2]

It is available in the ORD : 'data/7d/ord_dataset-7d8f5fd922d4497d91cb81489b052746.pb.gz'

Number of reactions	Number of variables
1536	Catalyst (16), Nucleophiles (16), Bases (6)



1.4 hte_dreher :

1.4.1 dreher_hte cheminformers library :

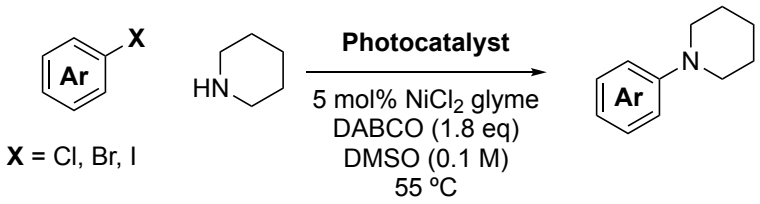
The Dreher dataset is a HTE dataset build in order to make cheminformers library for reaction benchmark, the data comes from experiment 2 presented in SI.[3]

The susbtrates are complex molecules of pharmaceutical interest (Ar-X, X = Cl, Br or I).

Not sure whether the SI experiment 1 is available or not but it is interesting as it is a kind of HTE optimization (48x32 = 1536 reactions) and exp. 2 is more like an HTE scope (1728 reactions).

It is available in the ORD : 'data/ac/ord_dataset-ac78456835404910b3a4c840248b6ac9.pb.gz'

Number of reactions	Number of variables
1728	Photocatalyst (11, amount :4), Ar-X (18), Piperidine (amount :2), Photocatalyst_2 (2 = one or nothing)



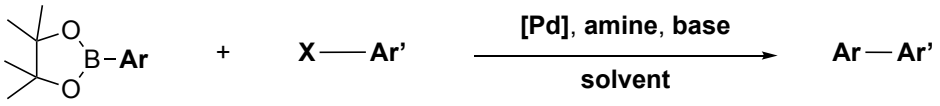
1.4.2 dreher_first cheminformers library :

The Dreher dataset is a HTE dataset build in order to make cheminformers library for reaction benchmark, the data comes from experiment 2 presented in SI.[4]

first 8 rows of Figure 4D

It is available in the ORD : 'data/ac/ord_dataset-ac78456835404910b3a4c840248b6ac9.pb.gz'

Number of reactions	Number of variables
264	boronate (24), aryl halides (18), Catalyst (8), Solvent (6), Bases(7), indole_halide (additive : 3), Base_2 (3)



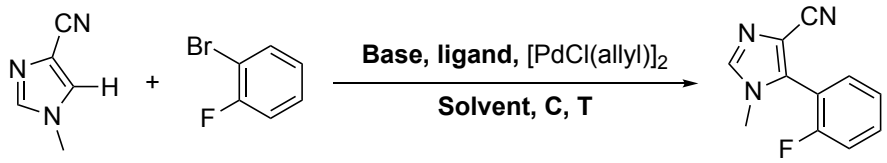
There are additionnal data on :
— cyanation methods of the pinacols boronates
— aryl halide CN coupling methods (Pd or Cu).

1.5 hte_shields :

The Shield HTE dataset is a direct arylation presented in order to benchmark the EDBO for reaction design.[8]
This dataset is constituted of one first full HTE dataset presented below and of an extended dataset constituted of the ful HTE screening for 2 new ligands, all others parameters varying (256/288 possible reactions performed).

It was downloaded from the EDBO git :https://github.com/b-shields/edbo

Number of reactions	Number of variables
1728	Ligands (12), Bases (4), Solvent (4), Concentrations(3) and Temperatures (3)

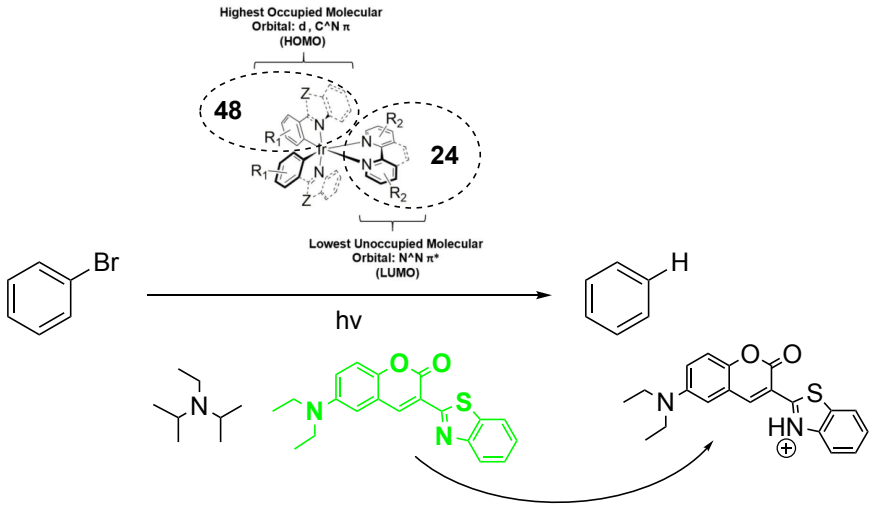


1.6 hte_photo :

The photo dataset is a HTE dataset build in order to screen Ir photocatalyst detailed below, for the dehalogenation of Ar-Br compounds. The output meausured is the conversion obtained after a certain time by UV-vis spectroscopy which considerably ease the analysis process.[5]
The only reaction parameter is the Ir(III) photo catalyst

It is available in the ORD : 'data/b4/ord_dataset-b440f8c90b6343189093770060fc4098.pb.gz'

Number of reactions	Number of variables
1152	Photocatalyst (1152 = 48x24)

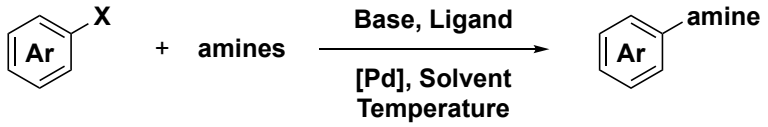


2 ELN :

The only ELN dataset available in the one of the preprint of Saebi.[7]
It is available in the ORD : 'data/00/ord_dataset-00005539a1e04c809a9a78647bea649c.pb.gz'

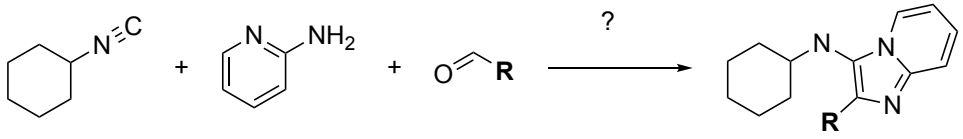
Number of reactions	Number of variables
750	Bases (13), Ligands (24), Catalyst (12), Ar-X (298), Amines (233), temperature (28), solvent (16)

For each reaction parameter the amount in moles and the solvent volume is detailed.



3 Imidazole :

A scope of multicomponent reactions. No reference given in the ORD.
It is available in the ORD : 'data/10/ord_dataset-10b940e7982c4622b1e1ac879394aba6.pb.gz'
384 aldéhydes for 384 reactions.

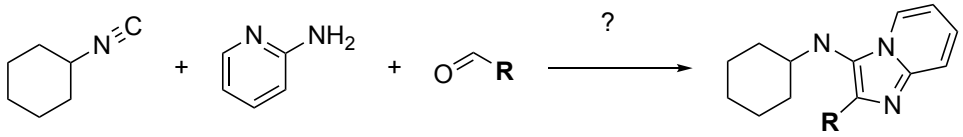


4 Suzuki Experiment :

Cliff activity for phosphines. (Newmann2021).
It is available in the ORD : 'data/3b/ord_dataset-3b5db90e337942ea886b8f5bc5e3aa72.pb.gz'

Number of reactions	Number of variables
450	Ar-Cl (2), ArB(OH)2 (5), ligands (90)

For each reaction parameter the amount in moles and the solvent volume is detailed.



Références

[1] Derek T Ahneman, Jesús G Estrada, Shishi Lin, Spencer D Dreher, and Abigail G Doyle. Predicting reaction performance in C–N cross-coupling using machine learning. *Science*, 360(6385) :186–190, apr 2018.

[2] Alexander Buitrago Santanilla, Erik L. Regalado, Tony Pereira, Michael Shevlin, Kevin Bateman, Louis-Charles Campeau, Jonathan Schneeweis, Simon Berritt, Zhi-Cai Shi, Philippe Nantermet, Yong Liu, Roy Helmy, Christopher J. Welch, Petr Vachal, Ian W. Davies, Tim Cernak, and Spencer D. Dreher. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science*, 347(6217) :49–53, jan 2015.

[3] Spencer D. Dreher and Shane W. Krska. Chemistry Informer Libraries : Conception, Early Experience, and Role in the Future of Cheminformatics. *Accounts of Chemical Research*, 54(7) :1586–1596, 2021.

[4] Peter S. Kutchukian, James F. Dropinski, Kevin D. Dykstra, Bing Li, Daniel A. Dirocco, Eric C. Streckfuss, Louis Charles Campeau, Tim Cernak, Petr Vachal, Ian W. Davies, Shane W. Krska, and Spencer D. Dreher. Chemistry informer libraries : A chemoinformatics enabled approach to evaluate and advance synthetic methods. *Chemical Science*, 7(4) :2604–2613, 2016.

[5] Velabo Mdluli, Stephen Diluzio, Jacqueline Lewis, Jakub F. Kowalewski, Timothy U. Connell, David Yaron, Tomasz Kowalewski, and Stefan Bernhard. High-throughput Synthesis and Screening of Iridium(III) Photocatalysts for the Fast and Chemoselective Dehalogenation of Aryl Bromides. *ACS Catalysis*, 10(13) :6977–6987, 2020.

[6] Damith Perera, Joseph W. Tucker, Shalini Brahmbhatt, Christopher J. Helal, Ashley Chong, William Farrell, Paul Richardson, and Neal W. Sach. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*, 359(6374) :429–434, jan 2018.

[7] Mandana Saebi, Bozhao Nan, John Herr, Jessica Wahlers, Zhichun Guo, Andrzej Zurański, Thierry Kogej, Per-Ola Norrby, Abigail Doyle, Olaf Wiest, and Nitesh Chawla. On the Use of Real-World Datasets for Reaction Yield Prediction. pages 1–24, 2021.

[8] Benjamin J. Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesus I. Martinez Alvarado, Jacob M. Janey, Ryan P. Adams, and Abigail G. Doyle. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844) :89–96, feb 2021.